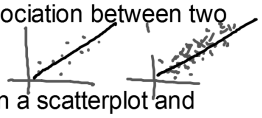
LinReg T-Test

$\bar{x}=9 \quad \mu=10$

* We want to test if there is an association between two **quantitative** variables (x and y)



* We look at a SAMPLE of data on a scatterplot and estimate the true relationship of the POPULATION plot

* LSR line (sample): $\hat{y} = b_0 + b_1 x$

Where: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$

* Population model (that we don't know):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

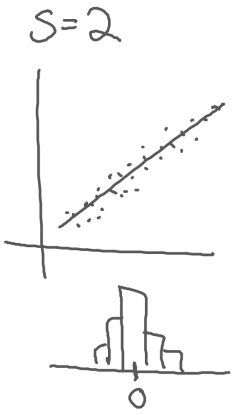
SAMPLE	POPULATION	WHAT IS IT?
b_0	β_0	intercept
b_1	β_1	slope
e_i	ε_i	errors (residuals)

Residuals (errors):

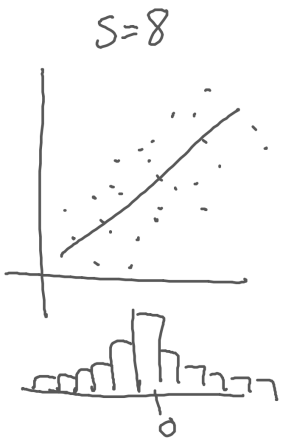
* Independent, Normally distributed
* N(0, S_e)

$s_e = \sqrt{\frac{\sum e^2}{n-2}} = df$ $\quad S = \sqrt{\frac{\sum (x_i-\bar{x})^2}{n-1}}$

small S_e of residuals:



large S_e of residuals:



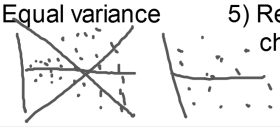
T-Test: Testing the slope of the population regression line

Hypotheses:

Ho: $\beta_1 = 0$	(there is no linear association btw X & Y)
Ha: $\beta_1 >, <, \neq 0$	(there is an association btw X & Y)

Conditions:

1) SRS	1) stated SRS or assumed represent.
2) Linear data	2) scatterplot is linear w/ no outliers
3) Independence	3) each piece of data can be assumed indep. of the others
4) Normal residuals	4) Normal prob. plot of residuals is linear (hist. unimodal, symm)
5) Equal variance	5) Residual plot shows no change in *spread* of residuals



Conditions met --> t-distribution --> LinReg t-test

Mechanics:

Test Statistic:

slope $\quad H_0: \beta_1 = 1$

t = statistic - parameter / std. dev. of stat (SE) $= \frac{b_1 - \beta_1}{SE_{b_1}} = \frac{b_1}{SE_{b_1}}$

P-Value: P(t $\gtrless$ test statistic) = tcdf(LB, UB, df)

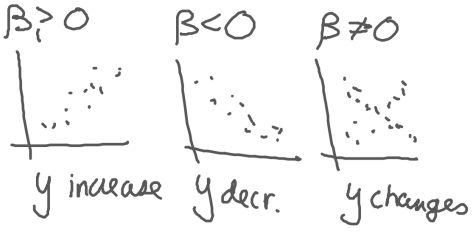
df = n - 2

Conclusion: (3 sentences)

* Reject / Fail to reject Ho ....

* We have sufficient/insufficient evidence that the slope of the population regression line is $>, <, \neq$ 0.

* Thus as X increases, Y _____.



Confidence Interval:

Formula: statistic ± (critical value)(SE) = (a, b)

$b_1 \pm t^* SE_{b_1}$

- Get the t* from the INVT program.

INVT
N=? 39
C=?
invT( )

- Be careful!
  * If sample size in the problem is 40, then df is n − 2 = 38
  * So in the INVT prgram you have to put in n = 39.
  * The program assumes that the df is n - 1.

Sentence:

We are ____% confident that as __X__ increases by 1 X unit, the __Y__ increases/decreases between __a__ and __b__ Y units.

Example: AIRFARES data
Assume we are told that a 95% confidence interval for the slope was (0.05435, 0.1804)

Interpret:
We are 95% confident that as the distance increases by 1 mile, the airfare increases between $0.05435 and $0.1804.

2 ways to do the mechanics of this test:

1) With actual data
   * Use LinReg t-test

Example: Airfare Data

WE WILL COME BACK TO THIS LATER

2) With computer output: $t = \frac{b_1 - \beta_1}{SE_{b_1}}$

Model of HusbandsAndWives — Multiple Regression

Response attribute (numeric): Age_Wife (Y)

Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	b_0 1.5740				
Age_Husband (X)	b_1 0.9112	0.0259	35.249	0.0000	0.8809

SE_{b_1} 0.124

Regression Equation: Age_Wife = 1.57400798388 + 0.911241590288 Age_Husband

R-Squared: **0.880894**
Adjusted R-Squared: **0.880186**
Standard Deviation of the Error: **3.95101**

$S_{resid} = S_e$

$n = 50$

$H_a: \neq$ (>)

$t = \frac{0.9112 - 0}{0.0259} = 35.181$ (35.249)

$P(t > 35.249) = tcdf(35.249, 1E99)$

NOTE: to check conditions, they will give you all the plots

$= 0$

Example:
1) Does a relationship exist between High School GPA and freshman year performance in college? A random sample of 40 freshman at a local college was taken and their HS GPA and the GPA from their first full year were recorded.

$n = 40$ $df = 38$

Hypotheses:

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Model of Sample of SATGPA — Multiple Regression

Response attribute (numeric): FYGPA

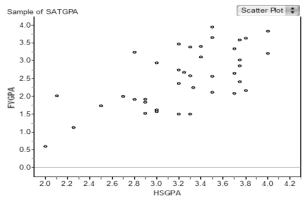
Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	b_0 1.0149				
HSGPA	b_1 1.0903	0.2045	5.333	0.0000	0.4280

SE

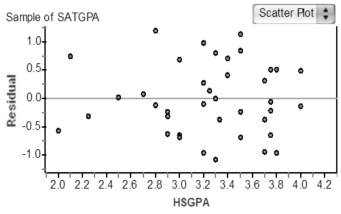
Regression Equation: FYGPA = − 1.01493878794 + 1.09034493208 HSGPA

R-Squared: **0.428036**
Adjusted R-Squared: **0.412985**
Standard Deviation of the Error: **0.626502**

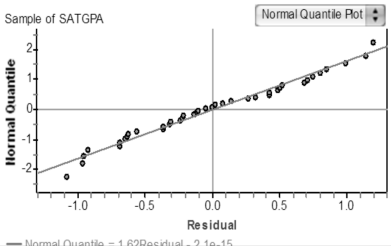
Scatterplot



Residual Plot



Normal Probability Plot



Conditions:

1) SRS	1) Stated random
2) Linear data	2) scatterplot of HSGPA vs. FYGPA is ≈ linear
3) Indep.	3) each college freshman is indep. of others
4) Normal resid.	4) Norm. prob. plot of resid. is ≈ linear
5) Equal spread	5) Resid. plot is scattered

Test Statistic:

Cond. met

t distrib

Linear Reg t test

$$t = \frac{b_1 - \beta_1}{SE} = \frac{1.0903}{0.2045} = 5.333$$

P-Value:

$$2 \cdot P(t > 5.333) = 0$$

Conclusion:

- Reject Ho b/c p-val < $\alpha = 0.05$.
- Suff. evid. that the slope of pop. regression line is not equal to 0.
- Concludes as HSGPA increases First year GPA changes.

Complete an appropriate confidence interval

$\neq$

$\alpha = 0.05$

95%

df = 38

$$b_1 \pm t^* SE$$

$$1.0903 \pm (2.024)(0.2045)$$

$$= (0.6764, 1.5042)$$

invT(0.025,38)

INVT

N=39

C=95

We are 95% conf. that as HSGPA increases by 1 pt, the first year GPA increases btw 0.6764 and 1.5042 pts.

Book problems: p. 674 #5, 13

⑤ Conditions:

1) SRS	1) assume that 2005 films are representative of all films
2) Linear data	2) roughly linear plot, one poss. outlier @ x = 185.
3) Independence	3) each movie is indep. of others
4) Normal residuals	4) Norm. prob. plot ≈ linear therefore normal resid. assume
5) Equal Variance	5) the resid. plot is scattered, no change in spread

cond. met → t distrib. → Lin. Reg. t test

$b_1 \pm t^*(SE_{b_1})$

$(0.7144) \pm (1.980)(0.1541)$

$(0.4093, 1.0195)$

df = 118

INVT
N = 119
C = 95%

We are 95% confident that for every increase of 1 min of run time of movie, the budget increases btw. 0.4093 and 1.0195 million dollars.

(13) a) $H_0: \beta_1 = 0$
$H_a: \beta_1 < 0$

df = 26

b) Histogram is right skewed, Normal residuals condition not met

c) Conditions not met → proceed anyway
t distrib → Lin Reg t test

$t = \frac{b_1 - \beta_1}{SE_b} = \frac{-0.02996}{0.0043} = -7.04$

$P(t < -7.04) = tcdf(-E99, -7.04, 26) = 8.907 \times 10^{-8}$

We reject Ho b/c p-value of 8.907×10^{-8} is less than $\alpha = 0.05$.

We have sufficient evidence that the slope of pop. regression line is less than 0.

Therefore as years increase, the diff. in the ages of 1st marriage of men & women decreases.

5) (a) Conditions:

1) SRS	1) assume 2005 movies are representative of all movies
2) Linear Data	2) The scatterplot looks roughly linear with 1 possible outlier
3) Independence	3) Each movie is indep. of others
4) Normal residuals	4) The normal probability plot of the residuals looks approx. linear --> normal residuals
5) Equal Variance	5) The residual plot shows no change in the spread of the residuals.

conditions met --> t distribution --> Lin Reg T-test

(b) $0.7144 \pm (1.98)(0.1541) = (0.4094, 1.0194)$

We are 95% confident that for every increase of 1 minute of run time of a movie, the budget increases btw 0.4094 and 1.0194 million dollars.

13) (a) $H_0: \beta_1 = 0$
$H_a: \beta_1 < 0$

(b) No, the conditions are not satisfied because the histogram is not normally distributed. It is right skewed

(c) $t = \frac{-0.02996}{0.0043} = -7.04$

$P(t < -7.04 \mid df = 26) = 8.907 \times 10^{-8}$

We reject Ho b/c pvalue of $8.907 \times 10^{-8} < \alpha = 0.05$.
We have sufficient evidence that the slope of the pop. regr. line between year and difference in marital ages of men & women is less than 0. Therefore as the year increases the difference between marital ages decreases.

p. 674
#4 (not b or e)
6
18